COUNTY HEALTH RANKINGS **WORKING PAPER**

SENSITIVITY ANALYSIS OF THE 2010 *COUNTY HEALTH RANKINGS*

Hyojun Park
Jessica K. Athens
Bridget C. Booske

JULY 2011

UNIVERSITY OF WISCONSIN
Population Health Institute
Translating Research into Policy and Practice

Table of Contents

Introduction....3

2010 County Health Rankings Methodology 3

Methods for Sensitivity Analysis 5

Analysis....8

Results: Overall Trends....8

Trends within States 11

Conclusion 12

References.... 13

Introduction

This report examines how changes in methodology affect the final Health Outcomes and Health Factors ranks in the 2010 *County Health Rankings* (the *Rankings*). We compared ranking results for all 50 states after applying alternative strategies to address two common data issues: (1) missing and unreliable values and (2) extreme (outlier) values. We also examined our assumptions regarding the relative contribution of each measure to overall Health Outcomes and Health Factors by testing different weighting schemes. Finally, we looked at scenarios in which alternative strategies for the three issues outlined above—missing or unreliable data; outlier values; and individual measure weights—were applied simultaneously.

2010 County Health Rankings Methodology

Overview

In the *Rankings,* ranks for Health Outcomes and Health Factors were assigned based on composite zscores of the measures that comprise these categories. Because the measures in the *Rankings* were not on the same scale, each variable was converted to a z-score, which represents the number of standard deviations an observation falls from the mean. Therefore, a z-score of 0 is equivalent to the mean of that measure across all counties in a state, negative z--scores represent values below the mean, and positive z-scores represent values above the mean. The majority of the measures in the *Rankings* are negative, so a higher value (positive z-score) is worse.[1]

Once the individual measures were converted to z-scores, they were used to compute overall Health Outcomes and Health Factors z-scores. Rather than average the z-scores of the individual measures, we applied a weighting scheme to represent the relative importance of a measure to Health Outcomes or Health Factors. As an example, the table below shows the weights of the measures that comprise Health Outcomes.

	Health Outcomes	**1.0**
MORTALITY	*Premature death*	0.5
MORBIDITY	*Self-reported health status*	0.1
	Physically unhealthy days	0.1
	Mentally unhealthy days	0.1
	Low birthweight births	0.2

The two constructs in Health Outcomes, mortality and morbidity, contributed equally to the overall Health Outcomes score. Among individual measures, premature death contributed the most to a county's performance in Health Outcomes, followed by low birthweight births, self-reported health, and unhealthy days.

After the composite scores for Health Outcomes and Health Factors were calculated, they were then ranked in ascending order, with the lowest (most negative) value ranking first (best) and the highest (most positive) value ranking last (worst).

[1] For six measures—primary care provider rate, diabetic screening, hospice use, high school graduation, college degrees, and access to healthy foods—a higher (positive) value is considered better. The z-scores for these measures were multiplied by -1 before calculating the composite z-scores.

Approaches to Problematic Values
Most measures in the 2010 *Rankings* had some problematic (missing, unreliable, or extreme) values due to sparsely populated counties that, for vital statistics data, either had no events or very few events over the observed time period. In the case of survey-based data, these counties may have had too few respondents to report an estimate. In an effort to reduce the number of counties with missing or unreliable data, we aggregated data over several years (3-7 years) for many of our measures; even so, some counties remained with few or no events; many survey-based estimates were also censored. In one data source, the American Community Survey's 3-Year Estimates—our key data source for social and economic measures—data for counties with populations less than 20,000 (N = 1,325 counties) were systematically censored.

A select list of counties with very poor data coverage across multiple measures were omitted from ranking (N = 126 counties). For the remaining counties, we applied the following strategies when faced with problematic values.

Missing values

- For missing values across all measures except those from the American Community Survey (ACS), we used a mean-substitution approach: where data were missing, we assigned a z-score of 0 for that measure.
- Because ACS data were (1) suppressed based on population size and (2) had equivalent Census 2000 measures, we decided to estimate missing values with a multiple imputation procedure that used population and Census estimates as instrumental variables. However, we only used multiple imputation if less than 50% of a state's counties were missing an ACS estimate. Otherwise, we reported the equivalent Census 2000 estimates for all counties in a state.

Unreliable values

- For vital statistics data—premature deaths, homicide, motor vehicle crash-related deaths, and low birthweight births—estimates were reported as long as 5 or more events were recorded over the time frame. However, if fewer than 20 events were recorded, the estimate was flagged as "unreliable" by the Centers for Disease Control and Prevention. We used unreliable estimates to calculate the z-scores, assuming that they were the best available estimates of these measures. We did not report any unreliable values.

Extreme values

- In the 2010 *Rankings,* we defined extreme or outlier values as those that were more than 3 standard deviations from the mean (a z-score of $|3.0|$). However, our primary concern was extreme values reported for smaller counties. Though large counties in some cases reported outlier values, we felt more confident that these represented the true values within those counties. To reduce the effect of extreme values that we suspected were statistical artifacts for counties with populations less than 20,000, we truncated individual measure z-scores that exceeded the $|3.0|$ threshold to -3.0 or +3.0.

Methods for Sensitivity Analysis

Testing Strategies for Missing and Unreliable Values
To test the impact of our strategies for handling missing and unreliable values, we considered five models that represent a combination of various approaches to missing and unreliable data:

Model A (reference): For missing values across all measures except ACS data, we employ mean substitution. For missing ACS data, we use multiple imputation to estimate missing values. We use the unreliable values as is. This approach is identical to that used in the 2010 *County Health Rankings*.

Model B: For missing values across all measures, including ACS data, we use mean substitution. Unreliable values were used as reported, except in the case of the mortality measure, in which we used mean substitution.

Model C: For both missing and unreliable values across all measures except ACS data, we use mean substitution. For missing ACS data, we use multiple imputation to estimate missing values.

Model D: No mean substitution is used for missing or unreliable values. For missing ACS data, we use multiple imputation to estimate missing values.

Model E: No mean substitution or imputation is used for missing or unreliable values, including ACS data.

Testing Strategies for Extreme Values
To test the impact of our strategy for extreme values, we compared three additional approaches to the 2010 strategy of truncating extreme values in order to stabilize measures and ranks.

Model I (reference): Z-scores beyond |3.0| are truncated only when a county's population size was less than 20,000. This approach is identical to that used in the 2010 *County Health Rankings*.

Model II: Z-scores beyond |3.0| are truncated regardless of population size.

Model III: Z-scores beyond |3.0| are truncated only when a county's standard error for a measure is greater than 20% of the point estimate.

Model IV: No truncation of z-scores is applied.

Testing Different Weighting Schemes
Weighting the effect of each measure is a critical component of the *County Health Rankings* methodology (for more detail, see [2] and [3]). The sensitivity analysis applies the weighting schemes from Booske *et al.* [3] to the *Rankings* data. See Table 1 for the exact weights applied in each model.

Model 1 (reference): Within Health Outcomes, the weights for the mortality measure (premature death) and morbidity measures (self-reported health, poor mental health days, poor physical health days, and low birthweight births) are set to 50:10:10:10:20. Therefore, each construct in Health Outcomes—mortality and morbidity—account for half the Health Outcomes composite z-score. Among Health Factors, the weights for the four constructs of Health Behaviors, Clinical Care, Social and Economic Factors, and Environment are set to 30:20:40:10. Each measure within a construct has a different weight. This approach is identical to that used in the 2010 *County Health Rankings*.

Model 2: Health Outcomes measures retain their weights from Model 1. Among Health Factors, the weights for Health Behaviors, Clinical Care, Social and Economic Factors, and Environment are set to 40:10:40:10. Each measure within a construct has a different weight. (This approach was based on a similar approach used for the *Wisconsin County Health Rankings* prior to 2010.)

Model 3: Even weights within Health Outcomes are applied. Similarly, the construct weights among Health Factors remain the same from Model 1 but each measure within a construct has the same weight.

Model 4: Even weights across Health Outcomes and Health Factors constructs are applied. Among Health Factors, weights for Health Behaviors, Clinical Care, Social and Economic Factors, and Environment are the same (25% each). Each measure within a construct has the same weight.

Model 5: Health Outcomes measures retain their weights from Model 1. The weight of each construct in Health Factors is determined based on regression analysis. Among Health Factors, weights for the four constructs of Health Behaviors, Clinical Care, Social and Economic Factors, and Environment are set to 25:19:52:3. Each measure within a construct has a different weight.

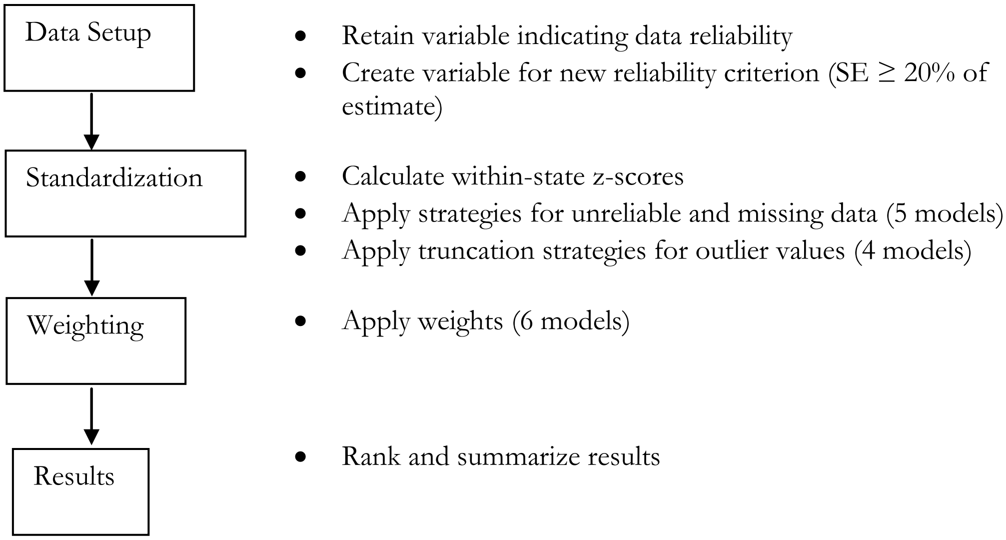
Model 6: Weights are assigned based on the weighting scheme use in *America's Health Rankings* (AHR), with slight modification due to differences in the measures used between AHR and the *County Health Rankings*. Among Health Factors, weights for the four constructs of Health Behaviors, Clinical Care, Social and Economic Factors, and Environment are set roughly to 27:37:30:7.

Table 1: Model Weights

Constructs	Measures	**Model 1** 2010 *Rankings*	**Model 2** 40:10:40:10 Health Factors	**Model 3** Even wgts. in construct	**Model 4** Even wgts. within & across constructs	**Model 5** Regression-based wgts.	**Model 6** AHR-based wgts.
MORTALITY	Premature death	0.500	0.500	0.200	0.500	0.500	0.300
MORBIDITY	Poor or fair health	0.100	0.100	0.200	0.100	0.100	0.200
	Poor physical health days	0.100	0.100	0.200	0.100	0.100	0.100
	Poor mental health days	0.100	0.100	0.200	0.100	0.100	0.100
	Low birthweight	0.200	0.200	0.200	0.200	0.200	0.300
HEALTH BEHAVIORS	Adult smoking	0.100	0.133	0.050	0.042	0.085	0.073
	Adult obesity	0.100	0.133	0.050	0.042	0.085	0.073
	Binge drinking	0.025	0.033	0.050	0.042	0.021	0.048
	Motor vehicle crash deaths	0.025	0.033	0.050	0.042	0.021	0.024
	Chlamydia rate	0.025	0.033	0.050	0.042	0.021	0.024
	Teen birth rate	0.025	0.033	0.050	0.042	0.021	0.024
CLINICAL CARE	Uninsured adults	0.050	0.025	0.040	0.050	0.048	0.073
	Primary care provider rate	0.050	0.025	0.040	0.050	0.048	0.073
	Preventable hospital stays	0.050	0.025	0.040	0.050	0.048	0.073
	Diabetic screening	0.025	0.013	0.040	0.050	0.024	0.073
	Hospice use	0.025	0.013	0.040	0.050	0.024	0.073
SOCIAL & ECONOMIC FACTORS	High school graduation	0.050	0.050	0.050	0.031	0.065	0.038
	College degrees	0.050	0.050	0.050	0.031	0.065	0.038
	Unemployment	0.100	0.100	0.050	0.031	0.130	0.038
	Children in poverty	0.075	0.075	0.050	0.031	0.098	0.038
	Income inequality	0.025	0.025	0.050	0.031	0.033	0.038
	Inadequate social support	0.025	0.025	0.050	0.031	0.033	0.038
	Single-parent households	0.025	0.025	0.050	0.031	0.033	0.038
	Violent crime/homicides	0.050	0.050	0.050	0.031	0.065	0.038
ENVIRONMENT	Particulate matter days	0.025	0.025	0.025	0.063	0.008	0.017
	Ozone days	0.025	0.025	0.025	0.063	0.008	0.017
	Access to healthy foods	0.025	0.025	0.025	0.063	0.008	0.017
	Liquor store density	0.025	0.025	0.025	0.063	0.008	0.017

Analysis

The figure below indicates how strategies for dealing with missing and unreliable data, extreme values, and changes in weights were applied to the *Rankings* data.



Given the number of models within each domain, we ultimately compared 120 different models. To examine the sensitivity of the *Ranking* results to changes in our approach to missing/unreliable values, outliers, and measure weights, we first examined how an isolated change in one domain—truncation, approaches to missing and unreliable values, or weights—affected the rank results. We calculated rank results for each strategy within domain, while applying the original *Rankings* strategy in the other two domains. These results were then correlated with the 2010 *Rankings* results to determine whether the changes had any effect.

Results: Overall Trends

None of the missing/unreliable strategies affected the rank results when applied without changes in the outlier strategies or weighting schemes (correlations between 2010 *Rankings* and all alternative models were 1.0). Similarly, none of the alternative outlier strategies resulted in any changes in the ranks.

The ranks were sensitive to changes in weights, but the correlations were still strong, indicating that the weights had a limited effect on the rank results. (For example, in West Virginia, a state with 55 counties, a strong rank-order correlation of 0.93 would result in most counties ranking in the same quartile under both weight schemes, with an average change of four ranks.) See table below for correlations; the two lowest correlation values for Health Outcomes and Health Factors are marked in bold.

	Model 2 40:10:40:10 Health Factors	**Model 3 Even wgts. in construct**	**Model 4 Even wgts. for all constructs**	**Model 5 Regression**	**Model 6 AHR**
Health Outcomes	1.00	**0.93**	1.00	1.00	**0.97**
Health Factors	0.99	0.96	**0.91**	0.99	**0.95**

Because Health Outcomes weights were only changed in models 3 and 6, these were the only models in which any variation was observed. Health Outcomes ranks changed most significantly with even weighting *within* constructs. Health Factors ranks were most sensitive to even weights *across* constructs. Nevertheless, the 2010 *Rankings* results appear to be quite robust with respect to changes in our strategies for missing and unreliable values, outlier values, or measure weights.

We then chose to examine which, if any, *combination* of alternative strategies across the three domains produced markedly different results from the 2010 *Rankings*. We generated correlation tables to represent these three domains; the values are the correlation across all U.S. counties between the reference model—the approach used in the 2010 *Rankings*—and the alternative strategies outlined previously. Considering the number of scenarios, we chose to generate four sets of tables, which should be read as followed:

- Each table represents a different truncation strategy for outlier values.
- The weighting schemes are shown in columns, labeled 1 through 6, and the strategies for missing and unreliable data are shown in rows, labeled A through E.
- The values in each cell represent the correlation between the reference model ranks and the ranks produced by the combined missing, outlier, and weighting strategies.
- The two lowest correlation values for Health Outcomes and Health Factors in each table are highlighted.

Outlier Strategy Model I (reference): Z-scores greater than |3.0| are truncated to +3.0 or -3.0 if the county's population is less than 20,000.

Missing/Unreliable Strategies		Weighting Schemes					
		1	2	3	4	5	6
Health Outcomes	A	1.00	1.00	0.93	1.00	1.00	0.97
	B	1.00	1.00	0.93	1.00	1.00	0.97
	C	1.00	1.00	**0.92**	1.00	1.00	0.97
	D	1.00	1.00	**0.90**	1.00	1.00	0.96
	E	1.00	1.00	**0.90**	1.00	1.00	0.96
Health Factors	A	1.00	0.99	0.96	**0.91**	0.99	0.95
	B	1.00	0.99	0.96	**0.91**	0.99	0.95
	C	1.00	0.99	0.96	**0.91**	0.99	0.95
	D	1.00	0.98	0.96	**0.91**	0.99	0.93
	E	1.00	0.98	0.94	**0.90**	0.99	0.93

Outlier Strategy Model II: Z-scores greater than |3.0| are truncated to +3.0 or -3.0 regardless of population size.

Missing/Unreliable Strategies		**Weighting Schemes** 1	2	3	4	5	6
Health Outcomes	A	1.00	1.00	0.93	1.00	1.00	0.97
	B	1.00	1.00	0.93	1.00	1.00	0.97
	C	1.00	1.00	**0.92**	1.00	1.00	0.97
	D	1.00	1.00	**0.90**	1.00	1.00	0.96
	E	1.00	1.00	**0.90**	1.00	1.00	0.96
Health Factors	A	1.00	0.99	0.96	**0.91**	0.99	0.95
	B	1.00	0.99	0.96	**0.91**	0.99	0.95
	C	1.00	0.99	0.96	**0.91**	0.99	0.95
	D	1.00	0.98	0.96	**0.91**	0.99	0.93
	E	1.00	0.98	0.94	**0.90**	0.99	0.93

Outlier Strategy Model III: Z-scores beyond |3.0| are truncated only when a county's standard error for a measure is greater than 20% of the point estimate.

Missing/Unreliable Strategies		**Weighting Schemes** 1	2	3	4	5	6
Health Outcomes	A	1.00	1.00	**0.93**	1.00	1.00	0.97
	B	1.00	1.00	**0.93**	1.00	1.00	0.97
	C	1.00	1.00	**0.92**	1.00	1.00	0.97
	D	1.00	1.00	**0.93**	1.00	1.00	0.97
	E	1.00	1.00	**0.93**	1.00	1.00	0.97
Health Factors	A	1.00	0.99	0.96	**0.91**	0.99	0.95
	B	1.00	0.99	0.96	**0.90**	0.99	0.95
	C	1.00	0.99	0.96	**0.91**	0.99	0.95
	D	1.00	0.99	0.96	**0.91**	0.99	0.95
	E	1.00	0.99	0.96	**0.90**	0.99	0.95

Outlier Strategy Model IV: No truncation of z-scores applied.

Missing/Unreliable Strategies		Weighting Schemes					
		1	2	3	4	5	6
Health Outcomes	A	1.00	1.00	**0.93**	1.00	1.00	0.97
	B	1.00	1.00	**0.93**	1.00	1.00	0.97
	C	1.00	1.00	**0.92**	1.00	1.00	0.97
	D	1.00	1.00	**0.93**	1.00	1.00	0.97
	E	1.00	1.00	**0.93**	1.00	1.00	0.97
Health Factors	A	1.00	0.99	0.96	**0.91**	0.99	0.95
	B	1.00	0.99	0.96	**0.90**	0.99	0.95
	C	1.00	0.99	0.96	**0.91**	0.99	0.95
	D	1.00	0.99	0.96	**0.91**	0.99	0.95
	E	1.00	0.99	0.96	**0.90**	0.99	0.95

For Health Outcomes, the correlations ranged between 0.90 and 1.00. In Health Factors, the lowest correlation across all scenarios was 0.90. Again, the rank results are most sensitive to changes in weights, following by changes in the strategy for missing and unreliable data. The outlier strategies seem to have minimal effect on the results. As noted earlier, the Health Outcomes ranks appeared most sensitive to even weighting within constructs (weight scheme 3). Health Factors ranks were again more reactive to changes in weights than Health Outcomes and changed most using weight scheme 4, even weights across constructs.

Trends within States

Though the ranks tend not to change dramatically overall, changes in methodology may result in significant in-state changes in rank, depending on a state's characteristics. We ran within-state rank-order correlations on z-scores between the reference model and the alternative models, and identified those instances in which a state had a correlation coefficient less than 0.8 (Cohen 1988). Not surprisingly, states that had low correlations across several models in both Health Outcomes and Health Factors had very few counties, such as Delaware (3 counties), Hawaii (5 counties), and Rhode Island (5 counties).

Within Health Outcomes, four states had correlations less than 0.8 in select models: Illinois, Iowa, Mississippi, and Nevada. All of the models that demonstrated lower correlations used even weights within the morbidity construct; for Illinois, Iowa, and Mississippi, the correlations were lowest in the models that used even weighting and did not apply mean substitution for missing or unreliable values. None of the truncation schemes affected the results in these states.

The unreliable values in these states tend to fall in the tails of the low birthweight data's distribution, as they are based on fewer observations; not using mean substitution for these estimates increases the number of extreme values that contribute to the composite morbidity score. Consequently, counties with unreliable values moved from the center of the distribution toward the tails. Treatment of unreliable values is particularly salient for Nevada, Iowa, and Illinois, which had unreliable low birthweight estimates for 30%, 20%, and 10% of their data, respectively.

Although the four morbidity measures—percent reporting fair or poor health, physically unhealthy days, mentally unhealthy days, and percent of births with low birthweight—tended to correlate between 0.32 and 0.75 nationally, these measures had lower rank-order correlations in the states of Illinois, Iowa, Mississippi, and Nevada. Mississippi and Nevada, particularly, demonstrated null or negative relationships between low birthweight and physically and mentally unhealthy days. As a result, the rankings were more strongly affected by changes in measure weights. In addition, choosing not to report mean values for missing data served as a *de facto* increase in weights for measures that had data reported.

Within Health Factors, Wyoming and West Virginia had correlations below 0.8 in all models that used even weighting (weight scheme 4) within and across Health Factor constructs. Neither truncation of z-scores nor strategies for missing and unreliable data had a significant effect on results. The increased weight accorded to the physical environment measures—2.5 times higher than the original 2010 *Rankings* model—contributed the most to the change in ranks. In West Virginia, the rank-order correlation between Health Factors composite scores using the original *Rankings* weights and the even weighting scheme was 0.77. However, when physical environment measures were not included in the composite score, the correlation was 0.92. Similarly, in Wyoming, the correlation between the reference model and the even weights model was 0.78 for Health Factors. Once physical environment measures were omitted, the rank results matched exactly.

Conclusion

The sensitivity analysis shows that, overall, the 2010 *County Health Rankings* results are robust even when different methodological strategies are applied. Across the three domains for which different strategies were tested, the weighting schemes, which represent the relative contribution of individual measures to overall performance in Health Outcomes and Health Factors, had the greatest effect on the results. Alternative strategies for missing/unreliable data and outlier values resulted in comparatively small changes in rank when compared to the reference model. Within-state ranks were more variable, dependent on the characteristics of that state.

References

1. University of Wisconsin Population Health Institute (UWPHI). County Health Rankings: Mobilizing Action Toward Community Health. UWPHI, 2010. Available from: http://www.countyhealthrankings.org/.

2. ———. County Health Rankings: Methods. UWPHI, 2010. Available from: http://www.countyhealthrankings.org/about-project/ranking-methods.

3. Booske BC, Athens JK, Kindig DA, et al. *Different Perspectives for Assigning Weights to Determinants of Health.* County Health Ranking Working Paper, UWPHI, 2010.

4. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* New York: Lawrence Erlbaum and Associates, 1988.